KINEMATICS-INDUCED MULTIMODAL 3D HUMAN POSE ESTIMATION WITH SUBJECT-LEVEL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal 3D Human Pose Estimation (3D HPE) combines complementary information from RGB, LiDAR, and mmWave radar, but models trained on correlated observations from the same individuals, raise privacy risks overlooked by record level analysis. We present a unified framework for multimodal 3D HPE that couples kinematics-induced sensor fusion with subject level privacy auditing and private training. First, our multimodal model aligns modality specific joint representation, injects skeletal structure and adaptively aggregates complementary sensor evidence for accurate pose prediction. Second, we formulate a black-box subject membership inference attack for 3D HPE, complemented by an empirical pointwise maximal leakage analysis, which characterizes how individual attack score outcomes change inference about the membership outcome. Third, we instantiate user-level differential privacy via Action Temporal Stratification, a population weighted within-subject sampling strategy that enforces action and temporal coverage. We evaluate our framework on the MM-Fi dataset across three diverse experimental protocols. Source-code will be released upon acceptance.

1 INTRODUCTION

Three dimensional human pose estimation (3D HPE) is a central problem in computer vision, supporting applications in healthcare monitoring, rehabilitation, sports analysis, and human computer interaction. The majority of existing approaches operate on RGB video, benefiting from large scale image datasets and mature two stage pipelines that detect 2D keypoints before lifting them into 3D space (Chen & Ramanan, 2017; Dang et al., 2023; Qi et al., 2019a; 2020b; 2019b; 2020a; Moon et al., 2019; Wehrbein et al., 2021; Pavlakos et al., 2017), with more recent methods incorporating temporal convolutions, transformers, kinematic priors, and diffusion based refinement (Pavllo et al., 2019; Zheng et al., 2021; Zhang et al., 2022; Zhu et al., 2023; Peng et al., 2024; Shan et al., 2023). However, RGB model deployment remains limited in unconstrained settings such as hospitals and assisted living environments due to performance degradation under occlusions, poor illumination, viewpoint variations and privacy-concerns. (Dang et al., 2023; Yang et al., 2024).

Because no single modality is reliable across all conditions, multimodal systems combining RGB, LiDAR, and millimeter-wave radar have emerged to leverage their complementary strengths (Fürst et al., 2021; An et al., 2022; Yang et al., 2023). However, effectively integrating these heterogeneousness modalities remains challenging. It requires resolving the differing feature distributions and noise characteristics of RGB, LiDAR, and radar into a common semantic space, while retaining the complementary, modality-specific evidence each sensor provides Chen & Yang (2025); Chae et al. (2024); Fürst et al. (2021); Yang et al. (2023).

A second, underexplored, concern in this setting is privacy. Human pose data captured through RGB, LiDAR, or radar can be linked back to individual subjects, yet the computer vision community has repeatedly been shown to overlook or mischaracterize this risk (Chen, 2019; Baldassini et al., 2025; Zhang & Zhao, 2025). Differential privacy (DP) provides a principled and mathematically grounded framework Dwork & Roth (2014) but applying it to visual prediction tasks via DP stochastic gradient descent (DP-SGD) incurs substantial utility loss (Golatkar et al., 2022; Zhou et al., 2021; Sivangi et al., 2026; Fang et al., 2023). Moreover, standard record-level DP protects individual samples but not the repeated participation of a subject across many records. This is especially problematic for LiDAR and radar-based human motion, whose four-dimensional spatio-temporal sequences are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

strongly correlated across frames from the same subject, so record-level guarantees can substantially understate an individual's true privacy risk (Ghazi et al., 2021; Chua et al., 2024).

Despite growing interest in both multimodal pose estimation and privacy preserving learning, the two problems have largely been studied in isolation. Audits of what a trained multimodal pose model reveals about its subjects remain scarce, with prior privacy risk analysis in this space typically framed as membership or attribute inference under record level assumptions (Gupta et al., 2021; Sivangi et al., 2026). This leaves open the question of how to design a multimodal pose estimation framework whose privacy is assessed and enforced at the level of the subject rather than the observation, and whether the resulting protection can be quantified beyond aggregate membership accuracy.

In this work, we present a unified framework that couples kinematics induced multimodal fusion with subject level privacy auditing and private training for 3D-HPE. Our contributions are threefold: **i)** First, we introduce a kinematics-induced fusion module featuring a **dual-track design with adaptive gating** that aligns modality-specific joint representations from RGB, LiDAR, and radar, and injects skeletal structure through fixed anatomical connectivity combined with learned joint dependencies. It adaptively aggregates the resulting evidence into a shared representation that outperforms each individual modality and achieves state-of-the-art performance. **ii)** Second, we formulate a black box subject level membership inference attack tailored to the regression nature of HPE and complement it with an empirical pointwise maximal leakage analysis **incorporating Dirichlet smoothing** (Saeidian et al., 2023), which characterizes how individual attack outcomes shift an adversary's belief about subject membership, offering a finer grained view of leakage than threshold based attack metrics alone. **iii)** Third, we instantiate differential privacy at the subject rather than the sample level, coupled with **private quantile-based adaptive clipping**, following the general user level formulation of Chua et al. (2024) and the correlated sampling perspective of Ghazi et al. (2021). Furthermore, we introduce **Action Temporal Stratification**, a population weighted within subject sampling strategy that preserves an unbiased subject gradient while enforcing coverage across actions and time.

2 Related Work

2.1 3D HPE using multimodal sensors

RGB based 3D HPE has become a central research direction in computer vision, largely driven by the availability of large scale image and video datasets and well established 2D pose detectors. Early approaches typically adopted a two-stage pipeline: first detecting 2D keypoints, then lifting these 2D poses into 3D spaceQi et al. (2019a; 2020b; 2019b; 2020a); Moon et al. (2019); Wehrbein et al. (2021); Pavlakos et al. (2017), with recent work using temporal convolutions, transformers, kinematic priors, and diffusion models to model spatial-temporal pose structure Pavllo et al. (2019); Zheng et al. (2021); Zhang et al. (2022); Zhu et al. (2023); Peng et al. (2024); Shan et al. (2023). Despite strong benchmark performance, RGB-based methods remain sensitive to occlusion, illumination, resolution, viewpoint changes, and scene clutter because they do not directly observe 3D geometry.

LiDAR and mmWave Radar have been explored as complementary sensing modalities for 3D HPE. LiDAR provides explicit 3D geometry and illumination invariance, enabling long-range markerless motion capture and robust 3D HPE from point clouds Li et al. (2022); Zhang et al. (2024); An et al. (2025); Zheng et al. (2022). However, LiDAR point clouds of humans are often sparse, incomplete, and sensitive to distance, self-occlusion, and sensor placement. mmWave Radar offers additional advantages under poor lighting and partial occlusion, and has been used for skeletal pose estimation, mesh recovery, diffusion-based pose estimation, and multi-view sensing Sengupta et al. (2020); Xue et al. (2021); Fan et al. (2024); Choi et al. (2025); Zhu et al. (2025). Yet Radar point clouds remain extremely sparse, noisy, and anatomically coarse compared with RGB and LiDAR observations.

Multimodal 3D HPE aims to combine these complementary cues for improved robustness and accuracy. Prior work has explored RGB-LiDAR pose estimation and multimodal human sensing datasets such as mRI and MM-Fi, which provide synchronized sensor streams for evaluating individual modalities and fusion strategies Fürst et al. (2021); An et al. (2022); Yang et al. (2023). However, most multimodal systems require multiple sensors during inference, increasing hardware cost, calibration complexity, synchronization overhead, and deployment constraints. This motivates teacher-student settings in which a multimodal teacher learns from complementary sensors during training, while a single-modality student performs efficient inference using only one available sensor.

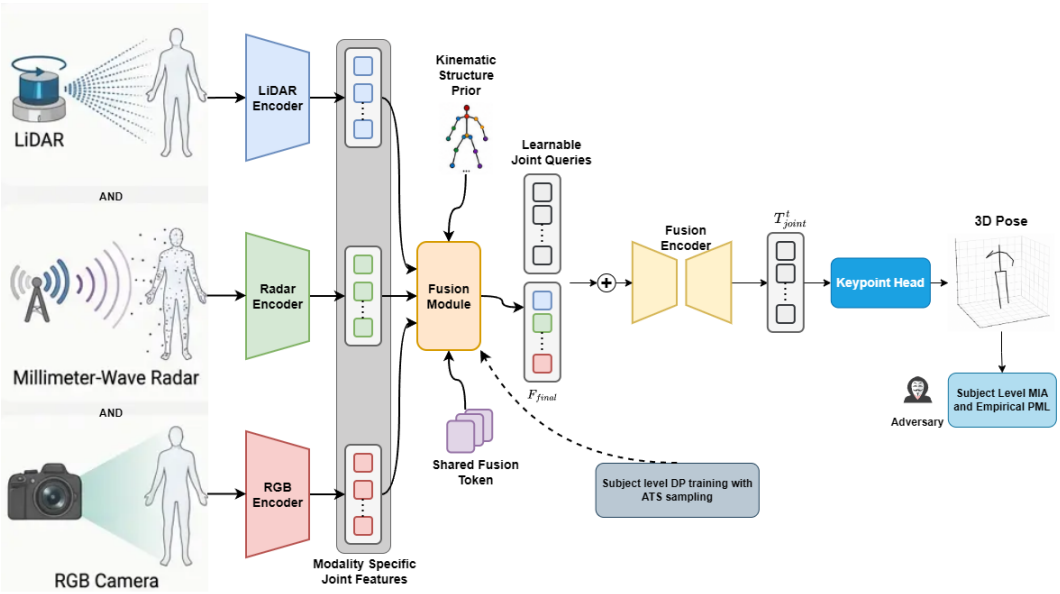

Figure 1: Overall framework of the proposed privacy aware multimodal 3D HPE method. RGB, LiDAR, and mmWave radar keypoint representations are integrated through the kinematics induced fusion module to predict the 3D pose. The model is trained on Action Temporal Stratified Subject level DP, while the final model is attacked through subject membership inference and empirical pointwise maximal leakage.

2.2 Privacy-preservation in computer vision

Growing evidence indicates that privacy preservation within computer vision community has frequently been ignored or misunderstood Zhang & Zhao (2025); Baldassini et al. (2025). The field has largely relied on fragmented, improvised anonymization techniques lacking formal guarantees. Rigorous alternatives such as differential privacy (DP) offer mathematical bounds, yet applying standard DP stochastic gradient descent to visual prediction tasks incurs a substantial utility loss Golatkar et al. (2022). Vision models suffer due to severe long-tail class imbalance that amplifies noise on minority classes, ie. rare poses, high dimensionality that reduces the signal-to-noise ratio of the gradient vectors Fang et al. (2023); Zhou et al. (2021), and continuous pixel values that create non-smooth loss landscapes violating Lipschitz continuity Fang et al. (2023). Consequently, DP-SGD performance relies critically on loss smoothness, dimensionality, and clipping calibration Lebensold et al. (2024); Sivangi et al. (2026).

Furthermore, it has been increasingly recognized that standard record level DP is insufficient when training data contains strong temporal correlations or multi sample contributions from individual subjects Ghazi et al. (2021); Chua et al. (2024); Kawamoto & Murakami (2019). This concern is particularly acute for LiDAR and Radar based human motion signatures, which are inherently four dimensional spatio-temporal point cloud sequences, with each subject contributing frames exhibiting substantial temporal autocorrelation. This paradigm requires user level, rather than sample level, privacy protection, using advanced mechanisms such as correlated sampling to handle structured per-user dependencies Ghazi et al. (2021).

3 Methodology

3.1 Overall Architecture

The proposed architecture employs a heterogeneous modality specific encoders followed by a kinematics induced multimodal fusion module for 3D HPE. We utilize heterogeneous backbones to extract modality-specific joint representations: a DSTFormer-based encoder for RGB derived 2D keypoint sequences, and PointTransformerV2-based encoders for LiDAR and Radar point clouds. The modality

specific representations are subsequently integrated through our kinematics induced fusion module for multimodal 3D HPE. This module uniquely maintains a modality specific joint representations together with shared fusion tokens, refining modality specific features through skeletal priors while simultaneously updating the shared representation through adaptive gating. This hierarchical design allows modality contributions to vary across anatomical joints and input samples through joint wise adaptive fusion weights. To preserve articulated structure, learnable joint queries are processed by a transformer-based keypoint head to produce one token for each of the $J = 17$ body joints Li et al. (2021). The multimodal model predicts the final 3D pose from these joint tokens using a regression head. The overall architecture is given in Figure 1.

3.2 Fusion Module

To effectively integrate LiDAR, Radar, and RGB modalities, we design a kinematics-induced fusion module that operates in three levels: modality alignment, kinematic-token fusion, and joint-wise aggregation. This architecture preserves modality-specific characteristics while progressively extracting a shared representation consistent across sensors and aware of skeletal constraints.

Modality Alignment: The three modalities exhibit different feature distributions due to their sensing mechanisms. LiDAR captures geometric structure, Radar offers sparse but robust measurements in any environment, and RGB provides rich visual and motion cues. Fusing these heterogeneous features can lead to unstable optimization and poor cross-modal agreement. We address this through a lightweight alignment stage that performs cross-modal calibration. Let $F_L, F_R, F_I \in \mathbb{R}^{J\times D}$ denote the features extracted from LiDAR, Radar and RGB derived keypoint backbone where J is the number of joints and D is the feature dimension. We append a learnable modality embedding e_m to each stream and apply multi-headed attention (MHA) jointly across the three modalities at each joint location which is given as:

$$\begin{aligned} \tilde{F_m} &= F_m + e_m, \quad m \in L, R, I \\ [X_L, X_R, X_I] &= [F_L, F_R, F_I] + MHA([\tilde{F_L}, \tilde{F_R}, \tilde{F_I}]) \end{aligned} \tag{1}$$

This interaction encourages the three streams to align at identical anatomical locations, allowing each modality to incorporate complementary evidence before the primary fusion stage.

Kinematic Injection and Dual Track Design: Post alignment, we employ a dual track design that maintains modality specific joint features alongside shared fusion tokens. The modality specific joint features retain local evidence from each sensor, while the shared tokens accumulate information that is jointly supported across all modalities. Within each stream, joint features are first refined via a kinematic propagation mechanism based on the human skeleton structure. Let $X_m \in \mathbb{R}^{J\times D}$ denote the modality specific joint features for modality m. Based on Peng et al. (2024), we define a modality dependent affinity matrix by combining the fixed skeletal adjacency A_{ske} with a learnable residual adjacency $\hat{A}_m$. This is then symmetrized to encode both known anatomical connectivity and learned long-range joint dependencies. We specifically update the kinematic refinement with the shared token which overall constitutes to calculating a gating coefficient. Formally,

$$\begin{aligned} A_m &= \frac{(A_{ske} + \hat{A}_m) + (A_{ske} + \hat{A}_m)^T}{2} \\ G_m &= \sigma(MLP([X_m; S])) \\ H_m &= X_m + A_m(G_m \odot X_m) \end{aligned} \tag{2}$$

where S is the initialized shared token ($S \in \mathbb{R}^{K\times D}$), G_m is an adaptive gating coefficient. This allows each modality to refine its joint representation in a structure-aware manner. Moreover, the gate is adjusted by current shared representation, implying the amount of kinematic refinement is not fixed and adapts according the cross-modal fusion state of the sample.

The modality specific features are then fused at each joint through adaptive weights. The model predicts per-joint fusion coefficients from the features themselves, allowing the contribution of each modality to vary across body parts and samples. This is given as:

$$F = \sum_m \alpha_m \odot H_m; \quad \sum_m \alpha_m = 1 \tag{3}$$

where the fusion weights α_m are produced by a lightweight MLP followed by softmax across modalities. This enables the contribution of each modality to adapt across anatomical joints and input samples.

The shared fusion tokens S are updated by attending to the fused representation

$$S' = S + MHA(S, F, F) \tag{4}$$

The shared tokens act as a cross-modal information pool. Since it receives information from fused anatomical representation, it gradually accumulates the consensus level information. The shared tokens are then fed back to each modality stream through cross-attention given as:

$$\begin{aligned} \Delta_m &= MHA(H_m, S', S') \\ X_m^{'} &= X_m + \Psi([H_m, F, \Delta_m]) \end{aligned} \tag{5}$$

where $\Psi(\cdot)$ denotes MLP layer. This ensures each of the modality-specific feature are more informative and cross-modally aware. We finally obtain the modality specific joint features $X_m^{'}$, together with shared token S'. A final joint-wise weighted fusion is then applied to produce the final fused representation:

$$F_{final} = \sum_{m \in L,R,I} \alpha_m^{final} \odot X'_m \tag{6}$$

This final representation summarizes the complementary strengths of the three modalities while preserving the anatomical consistency through kinematic refinement.

3.3 Action Temporal Stratified Subject-Level Differential Privacy

Multi-modal human pose dataset contain many observations contributed by the same subject, including temporally overlapping windows within individual actions. A record level DP mechanism that treats each temporal window as an independent privacy unit protects the inclusion or removal of one window, but does not directly protect the inclusion or removal of a subject's complete contribution. Towards this, we instantiate subject level differential privacy where we treat each subject's entire records as the privacy unit following the general user wise formulation of Chua et al. (2024) and prior user level private training with multiple samples per user Ghazi et al. (2021). For our setting, all the multi-modal pose windows belonging to one subject constitute a single privacy unit.

Subject-level Neighborhood: Let the training dataset consist of N subjects,

$$\mathcal{D} = \biguplus_{s=1}^{N} \mathcal{D}_s, \qquad \mathcal{D}_s = \{z_{s,1}, \ldots, z_{s,M_s}\}, \tag{7}$$

where $z_{s,i}$ denotes one multimodal pose window and M_s is the number of available windows for subject s. Two datasets are neighbors if they differ by the complete contribution of one subject.

$$\mathcal{D}' = \mathcal{D} \setminus \mathcal{D}_s. \tag{8}$$

Differential Privacy: A randomized training mechanism $\mathcal{M}$ satisfies subject-level (ϵ, δ)-differential privacy if, for every neighboring $\mathcal{D}, \mathcal{D}'$ and measurable output set $\mathcal{O}$,

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{O}] + \delta. \tag{9}$$

This is the standard differential privacy definition Dwork & Roth (2014), with neighborhood specialized to complete subject removal rather than removal of a single pose window.

Subject-wise private gradient: At private optimization step t, each of the N subject slots is included independently with probability ρ,

$$I_{t,s} \sim \text{Bernoulli}(\rho), \qquad \mathcal{S}_t = \{s : I_{t,s} = 1\}. \tag{10}$$

This constitutes Poisson sampling at the protected subject level. For each sampled subject, K pose windows are used to construct one subject-gradient estimate $g_{t,s}$ and clipping is applied only after this within subject aggregation Chua et al. (2024), where multiple records belonging to a sampled user are first averaged into one user contribution.

The resulting subject vector is clipped once with a global ℓ_2 bound C_t,

$$\bar{g}_{t,s} = g_{t,s} \min\left(1, \frac{C_t}{\|g_{t,s}\|_2}\right), \qquad \|\bar{g}_{t,s}\|_2 \le C_t. \tag{11}$$

The privatized aggregate is

$$\widetilde{g}_t = \frac{1}{\rho N}\left[\sum_{s\in\mathcal{S}_t} \bar{g}_{t,s} + \sigma_g C_t \xi_t\right], \qquad \xi_t \sim \mathcal{N}(0, I), \tag{12}$$

where ρN is the public expected number of sampled subjects. Since subsequent optimizer operations depend only on privatized gradients, they are post-processing of the private mechanism Dwork & Roth (2014). To avoid manual tuning of this hyperparameter, we adapt the private quantile-based adaptive clipping mechanism from Andrew et al. (2021) to complete subject gradients, as detailed in Appendix A.

Why stratify windows within a subject? Uniform within-subject sampling provides an unbiased estimate of a subject's mean gradient, but pose sequences exhibit substantial semantic and temporal redundancy. With a finite record budget K, a random draw may therefore over-represent a frequently occurring action or a short temporal region while missing other behaviors of the same subject. Such heterogeneous sampling can increase stochastic-gradient variance, motivating stratified sampling in stochastic optimization Zhao & Zhang (2014); Lu et al. (2021).

For fixed parameters θ, let the complete pre-clipping gradient of subject s be

$$\mu_s(\theta) = \frac{1}{M_s}\sum_{z\in\mathcal{D}_s} \nabla_\theta \ell(\theta; z). \tag{13}$$

Our goal is to approximate $\mu_s(\theta)$ using only $K \ll M_s$ windows without changing the intended complete subject mean gradient. We therefore partition $\mathcal{D}_s$ by action and temporal position and construct the population-weighted estimator

$$g_s^{\text{ATS}} = \sum_{h=1}^{H_s} \frac{M_{s,h}}{M_s} g_{s,h}, \tag{14}$$

where $M_{s,h}$ is the population of stratum h and $g_{s,h}$ is its sampled mean gradient. As proven in Appendix A (Lemma 1), population weighting ensures an unbiased pre-clipping subject gradient in expectation, enforcing robust action and temporal coverage without altering the true subject mean. In particular,

$$\mathbb{E}\big[g_s^{\text{ATS}} \mid \mathcal{D}_s, \theta\big] = \mu_s(\theta), \tag{15}$$

Thus, ATS changes how the within subject sampling budget is allocated, rather than changing the intended pre clipping gradient in expectation. A formal derivation of this result is provided in Appendix A.

Action Temporal Stratification (ATS): For subject s, we partition the legal windows into disjoint and exhaustive strata,

$$\mathcal{D}_s = \biguplus_{h=1}^{H_s} \mathcal{D}_{s,h}, \qquad M_{s,h} = |\mathcal{D}_{s,h}|, \qquad \sum_{h=1}^{H_s} M_{s,h} = M_s. \tag{16}$$

Strata are constructed using action identity and chronological position within each action. In practice, we first allocate sampling slots across non-empty actions and subsequently divide each action sequence into contiguous temporal bins.

Let A_s be the number of non-empty actions for subject s. We reserve one sampling slot for every action and distribute the remaining $R_s = K - A_s$ slots approximately in proportion to proportion to the number of windows in each action. For an action a containing $M_{s,a}$ windows, define

$$q_{s,a} = R_s \frac{M_{s,a}}{M_s}, \qquad n_{s,a} = 1 + \lfloor q_{s,a} \rfloor + \Delta_{s,a}, \tag{17}$$

where $\Delta_{s,a} \in \{0, 1\}$ is assigned to actions with the largest fractional remainders such that $\sum_a n_{s,a} = K$. Windows within each action are sorted chronologically and divided into contiguous temporal bins according to allocated number of slots $n_{s,a}$.

More generally, suppose $n_{s,h} \geq 1$ samples are drawn uniformly from stratum h, with

$$\sum_{h=1}^{H_s} n_{s,h} = K. \tag{18}$$

The mean gradient within the sampled stratum is

$$g_{s,h} = \frac{1}{n_{s,h}} \sum_{j=1}^{n_{s,h}} \nabla_\theta \ell(\theta; Z_{s,h,j}), \tag{19}$$

where $Z_{s,h,j}$ is sampled uniformly from $\mathcal{D}_{s,h}$. Our Action Temporal Stratified estimator is

$$g_s^{\mathrm{ATS}} = \sum_{h=1}^{H_s} \frac{M_{s,h}}{M_s} g_{s,h}. \tag{20}$$

Equivalently, every selected draw from stratum h receives weight

$$w_{s,h,j} = \frac{M_{s,h}}{M_s n_{s,h}}. \tag{21}$$

Population weighting allows us to enforce action and temporal coverage without assigning disproportionate objective weight to smaller strata. Importantly, ATS modifies only the construction of the pre-clipping subject gradient. The resulting vector is still clipped once at the subject level as in Eq.11. Consequently, its contribution remains bounded by C_t, while the outer Bernoulli subject sampling and Gaussian mechanism remain unchanged. Appendix A (Proposition 1) formally establishes that replacing uniform sampling with ATS does not alter the outer subject-level privacy accounting or require modifications to our Rényi differential privacy composition. For fixed ρ, effective noise multiplier σ, total private steps T,δ, and adaptive clipping configuration, ATS uses the same subject level privacy accounting as random within subject sampling. We provide the formal privacy argument in Appendix A. We adapt private adaptive subject level clipping following Andrew et al. (2021); σ_g in Eq.12 denotes the resulting gradient-noise multiplier after jointly accounting for the private clipping-count query. We use Rényi differential privacy to compose the repeated sampled gaussian mechanisms Mironov (2017). The complete adaptive-clipping derivation, noise calibration are provided in Appendix A.

3.4 Subject Level Membership Inference and Pointwise Maximal Leakage

We evaluate membership evidence using the same subject-level units from private training. While membership inference measures whether an adversary can distinguish training from non training subjects, pointwise maximal leakage (PML) characterizes how individual score outcomes shift membership beliefs.

Threat model and membership experiment: Following the privacy game formulation of Salem et al. (2023) and the subject membership setting from Suri et al. (2022), let $\mathcal{S}_{\mathrm{tr}}$ denote the subjects included in the main training pool and define indicator as

$$B_s = \mathbf{1}\{s \in \mathcal{S}_{\mathrm{tr}}\}. \tag{22}$$

Membership only denotes inclusion in the training dataset. The attacker has prediction only access to the fixed final model f_{θ_T} and Q query windows from the target subject, together with the 3D pose annotations. The adversary does not have access to model parameters, gradients, intermediate features, or metadata that would reveal any membership. This is a ground truth aware black-box threat model. Membership labels are defined externally by the experiment split and are never obtained from the trained checkpoint.

Subject level membership score: As 3D HPE is a regression problem, we use prediction error as the membership inference signalYeom et al. (2018); Gupta et al. (2021). For each query window i, we compute the MPJPE error e_i to derive the subject-level membership inference score S_Q:

$$S_Q(s) = -\frac{1}{Q}\sum_{i=1}^{Q} e_i \tag{23}$$

For each subject, the Q queried windows produce a score where a higher value corresponds to a lower pose error and stronger membership evidence under this attack. We use subject level ROC-AUC as the primary threshold independent metric. For threshold-based evaluation, we define $\hat{B}_s = 1\{S_Q(s) \geq \tau\}$, where τ is selected by stratified subject level cross fitting using balanced accuracy on remaining folds.

Empirical Pointwise Maximal Leakage: While MIA measures how well the attack scores distinguish about the membership class, PML measures the increase in inference confidence of the adversary associated with that particular outcome. We define the audit game with a balanced membership prior as $P(B = 0) = P(B = 1) = 1/2$ Saeidian et al. (2023). Since S_Q is continuous, we use a small set of fixed score boundaries, and this is declared independently of the evaluated scores, to obtain the discrete evidence variable $Y \in \{0, \ldots, H-1\}$. Let $n_{b,h}$ be the number of subjects with membership label b and let score lie in bin h, and $N_b = \sum_h n_{b,h}$. To prevent zero probability estimates in small subject setting, we use Dirichlet-1/2 smoothing:

$$\hat{p}_b(h) = \frac{n_{b,h} + \frac{1}{2}}{N_b + \frac{H}{2}}. \tag{24}$$

For the balanced binary membership bit, the PMLSaeidian et al. (2023) definition specializes as:

$$\ell(B \to h) = \log \frac{2\max\{\hat{p}_0(h), \hat{p}_1(h)\}}{\hat{p}_0(h) + \hat{p}_1(h)}, \qquad 0 \leq \ell(B \to h) \leq \log 2. \tag{25}$$

A value of zero denotes that the observed outcome does not change the balanced membership belief, whereas $log2$ corresponds to complete disclosure of the membership variable. As smoothing assigns positive probability even to bins unobserved in both classes, the primary empirical summary takes the maximum only over bins containing at least one observed subject.

$$\widehat{\ell}_{\max}^{obs} = \max_{h: n_{0,h}+n_{1,h}>0} \ell(B \to h). \tag{26}$$

This is empirical estimate for the specified MIA and is not to be interpreted as a formal guarantee.

4 Experimental Results

4.1 Results

4.1.1 Effect of Multimodal Fusion

The main experimental setup is described in Appendix B. As shown in Table 1, under protocol 3, which considers all activities, the proposed fusion model achieves the lowest MPJPE scores across all three evaluation splits. Compared with the strongest unimodal implementation in each split, fusion reduces MPJPE by 5.46mm, under random split, 10.81mm under cross environment split and 11.05mm under cross subject split. The larger improvements under cross environment and cross subject settings suggest that combining complementary information from RGB, LiDAR, and Radar is particularly beneficial under distribution shifts. Results for protocols 1 and 2 are provided in Appendix B.2, while additional modality ablations and comparisons with a BeV based fusion model are presented in Appendix B.3, B.4 respectively.

4.1.2 ATS Subject Level DP and Attack inference

Table 2 demonstrates the subject-level privacy-utility tradeoff. Stronger DP suppresses membership evidence: $\epsilon = 1$ yields near chance ROC-AUC and zero empirical PML at a sever utility cost. As ϵ

Model	Modality	Random		Cross-Environment		Cross-Subject	
		MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Protocol 3: All Activities							
Existing Methods							
MM-Fi Yang et al. (2023)	RGB	279.0	81.2	288.6	84.1	285.3	81.9
MM-Fi Yang et al. (2023)	LiDAR	92.5	61.5	303.8	133.0	103.8	61.5
MM-Fi Yang et al. (2023)	Radar	117.0	57.3	161.6	73.7	129.7	60.0
MM-Fi Yang et al. (2023)	RGB+LiDAR	87.5	59.8	192.1	89.6	104.6	62.1
X-Fi Chen & Yang (2025)	RGB	93.9	60.3	–	–	–	–
X-Fi Chen & Yang (2025)	LiDAR	167.1	103.2	–	–	–	–
X-Fi Chen & Yang (2025)	Radar	127.4	69.8	–	–	–	–
X-Fi Chen & Yang (2025)	RGB+LiDAR	93.0	59.7	–	–	–	–
X-Fi Chen & Yang (2025)	RGB+Radar	88.8	57.3	–	–	–	–
X-Fi Chen & Yang (2025)	LiDAR+Radar	109.8	63.4	–	–	–	–
X-Fi Chen & Yang (2025)	RGB+LiDAR+Radar	88.4	57.2	–	–	–	–
Our Implementations							
PointTransformerV2	LiDAR	46.24	36.13	60.70	40.12	55.28	38.21
PointTransformerV2	Radar	61.09	45.27	78.42	58.58	61.80	47.11
DSTFormer	RGB	44.34	32.84	63.30	48.03	48.70	36.87
Fusion	RGB+LiDAR+Radar	**38.88**	29.45	**49.89**	40.03	**37.65**	26.04

Table 1: Comparison on Protocol 3 using all activities. Lower MPJPE and PA-MPJPE indicate better performance.

ϵ	MPJPE (mm) ↓	ROC-AUC	Empirical PML (nats)
1	1081.32	0.48	0.0
16	314.85	0.59	0.14
64	278.58	0.64	0.14
100	269.86	0.65	0.29
200	247.72	0.73	0.29
∞	37.65	0.83	0.44

Table 2: Privacy–utility and empirical leakage results under the MM-Fi Protocol 3 cross-subject setting. Each row corresponds to a subject-level differentially private model trained at the indicated privacy budget ϵ and non-dp baseline is also provided. MPJPE measures 3D pose estimation utility, ROC-AUC measures subject-membership discrimination, and empirical PML characterizes leakage in the corresponding membership-evidence channel. Lower MPJPE and PML are better, while ROC-AUC closer to 0.5 indicates weaker membership distinguishability.

increases pose accuracy improves alongside attack distinguishability. The non-private model ($\epsilon = \infty$) achieves the best MPJPE but also the strongest membership signal, with 7 out of the 8 held out subjects falling in the higher error half of the attack scores. With only 8 subjects held-out, only the non-private model and $\epsilon = 200$ are individually distinguishable from chance. PML changes more coarsely than ROC-AUC as it relies on discretized score outcomes.

5 Conclusion

In this work we present a framework for multimodal 3D human pose estimation that combines kinematics-induced fusion with subject-level DP training, and empirical membership inference. Overall, our fusion model improves the pose accuracy over individual modalities. The proposed Action Temporal Stratification constructs population-weighted subject gradients, while membership inference and PML provide complementary evidence of model privacy. The results show utility degradation under subject-level DP, particularly at stricter privacy budgets. Our evaluation is limited to the predefined MM-Fi split of subjects with eight non-member subjects. Future work should investigate more effective private adaptation techniques and evaluate on broader subject cohorts.

AI USE STATEMENT

Generative AI tools were used solely to improve the writing (grammar, clarity and readability of the manuscript). They are not used to develop research methodology, formulate hypotheses, design experiments or any science claims. AI assisted edits are reviewed by the authors and take responsibility for the final manuscript.

REFERENCES

Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in neural information processing systems*, 35: 27414–27426, 2022.

Xiaoqi An, Lin Zhao, Chen Gong, Jun Li, and Jian Yang. Pre-training a density-aware pose transformer for robust lidar-based 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1755–1763, 2025.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in neural information processing systems*, 34:17455–17466, 2021.

Michele Baldassini, Francesco Pistolesi, and Beatrice Lazzerini. Don't call it privacy-preserving or human-centric pose estimation if you don't measure privacy. In *NeurIPS 2025 Position Paper Track*, 2025.

Yujeong Chae, Hyeonseong Kim, and Kuk-Jin Yoon. Towards Robust 3D Object Detection with LiDAR and 4D Radar Fusion in Various Weather Conditions . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15162–15172, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01436. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01436.

Ching-Hang Chen and Deva Ramanan. 3D Human Pose Estimation = 2D Pose Estimation + Matching . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5759–5767, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.610. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.610.

Li Chen. Understanding the efficacy, reliability and resiliency of computer vision techniques for malware detection and future research directions. *arXiv preprint arXiv:1904.10504*, 2019.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, volume 2025, pp. 97378–97397, 2025.

Jaeho Choi, Soheil Hor, Shubo Yang, and Amin Arbabian. Mvdoppler-pose: Multi-modal multi-view mmwave sensing for long-distance self-occluded human walking pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27750–27759, 2025.

Lynn Chua, Badih Ghazi, Yangsibo Huang, Pritish Kamath, Ravi Kumar, Daogao Liu, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Mind the privacy unit! user-level differential privacy for language model fine-tuning. *arXiv preprint arXiv:2406.14322*, 2024.

Bowen Dang, Xi Zhao, Bowen Zhang, and He Wang. Reconstructing 3d human pose from rgb-d data with occlusions. In *Computer Graphics Forum*, volume 42, pp. e14982. Wiley Online Library, 2023.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Junqiao Fan, Jianfei Yang, Yuecong Xu, and Lihua Xie. Diffusion model is a good pose estimator from 3d rf-vision. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. Improved convergence of differential private SGD with gradient clipping. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FRLswckPXQ5.

Michael Fürst, Shriya TP Gupta, René Schuster, Oliver Wasenmüller, and Didier Stricker. Hperl: 3d human pose estimation from rgb and lidar. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7321–7327. IEEE, 2021.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. *Advances in Neural Information Processing Systems*, 34:20172–20184, 2021.

Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed differential privacy in computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051, 2023.

Umang Gupta, Dimitris Stripelis, Pradeep K Lam, Paul Thompson, Jose Luis Ambite, and Greg Ver Steeg. Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*, pp. 228–251. PMLR, 2021.

Yuzhe Ji, Yijie Chen, Liuqing Yang, Rui Ding, Meng Yang, and Xinhu Zheng. Vexkd: The versatile integration of cross-modal fusion and knowledge distillation for 3d perception. *Advances in Neural Information Processing Systems*, 37:125608–125634, 2024.

Yusuke Kawamoto and Takao Murakami. Local obfuscation mechanisms for hiding probability distributions. In Kazue Sako, Steve Schneider, and Peter Y. A. Ryan (eds.), *Computer Security – ESORICS 2019*, 2019.

Jonathan Lebensold, Doina Precup, and Borja Balle. On the privacy of selection mechanisms with gaussian noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 1495–1503. PMLR, 2024.

Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20502–20512, 2022.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11293–11302, 2021. doi: 10.1109/ICCV48922.2021.01112.

Yucheng Lu, Youngsuk Park, Lifan Chen, Yuyang Wang, Christopher De Sa, and Dean Foster. Variance reduced training with stratified sampling for forecasting models. In *International Conference on Machine Learning*, pp. 7145–7155. PMLR, 2021.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10133–10142, 2019.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7025–7034, 2017.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7753–7762, 2019.

Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1123–1132, 2024.

Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3957–3966, 2019a.

Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8):2617–2633, 2019b.

Mengshi Qi, Jie Qin, Yu Wu, and Yi Yang. Imitative non-autoregressive modeling for trajectory forecasting and imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12736–12745, 2020a.

Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Stc-gan: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. *IEEE Transactions on Image Processing*, 29: 5420–5430, 2020b.

Mengshi Qi, Jiaxuan Peng, Xianlin Zhang, and Huadong Ma. Towards balanced multi-modal learning in 3d human pose estimation. *arXiv preprint arXiv:2501.05264*, 2025.

Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Pointwise maximal leakage. *IEEE Transactions on Information Theory*, 2023.

Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 327–345. IEEE, 2023.

Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE sensors journal*, 20(17):10032–10044, 2020.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14761–14771, 2023.

Kaushik Bhargav Sivangi, Paul Henderson, and Fani Deligianni. Differentially private 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Denver, Colorado, USA, 2026.

Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317*, 2022.

Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11199–11208, 2021.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.

Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th annual international conference on mobile systems, applications, and services*, pp. 269–282, 2021.

Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023.

Yanni Yang, Pengfei Hu, Jiaxing Shen, Haiming Cheng, Zhenlin An, and Xiulong Liu. Privacy-preserving human activity sensing: A survey. *High-Confidence Computing*, 4(1):100204, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Jingyi Zhang, Qihong Mao, Siqi Shen, Chenglu Wen, Lan Xu, and Cheng Wang. Lidarcapv2: 3d human pose estimation with human–object interaction from lidar point clouds. *Pattern Recognition*, 156:110848, 2024.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13232–13242, 2022.

Xingchen Zhang and Zixian Zhao. More effort is needed to protect pedestrian privacy in the era of ai. In *NeurIPS 2025 Position Paper Track*, 2025.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Peilin Zhao and Tong Zhang. Accelerating minibatch stochastic gradient descent using stratified sampling. *arXiv preprint arXiv:1405.3080*, 2014.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11656–11665, 2021.

Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4478–4487, 2022.

Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private {sgd} with gradient subspace identification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7dpmlkBuJFC.

Bing Zhu, Zixin He, Weiyi Xiong, Guanhua Ding, Tao Huang, and Wei Xiang. Probradarm3f: mmwave radar-based human skeletal pose estimation with probability map guided multi-format feature fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 2025.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15085–15099, 2023.

A Appendix

Lemma 1: Unbiased pre-clipping subject gradient. Conditioned on $\mathcal{D}_s$ and θ, the ATS estimator satisfies

$$\mathbb{E}\left[g_s^{\mathrm{ATS}}\right] = \mu_s(\theta). \tag{27}$$

Proof. Uniform sampling within stratum h gives

$$\mathbb{E}[g_{s,h}] = \frac{1}{M_{s,h}} \sum_{z \in \mathcal{D}_{s,h}} \nabla_\theta \ell(\theta; z). \tag{28}$$

Therefore,

$$\mathbb{E}[g_s^{\mathrm{ATS}}] = \sum_{h=1}^{H_s} \frac{M_{s,h}}{M_s} \mathbb{E}[g_{s,h}] \tag{29}$$

$$= \frac{1}{M_s} \sum_{h=1}^{H_s} \sum_{z \in \mathcal{D}_{s,h}} \nabla_\theta \ell(\theta; z) \tag{30}$$

$$= \frac{1}{M_s} \sum_{z \in \mathcal{D}_s} \nabla_\theta \ell(\theta; z) = \mu_s(\theta). \tag{31}$$

The argument requires uniform marginals within each stratum but does not require independence between records sampled from the same stratum. □

Private adaptive subject clipping. A fixed clipping threshold can either excessively truncate useful subject gradients or induce unnecessarily large absolute Gaussian noise. We therefore adapt the private quantile-based clipping mechanism of Andrew et al. (2021) from user updates to complete subject gradients. For each sampled subject, define the centered clipping indicator

$$a_{t,s} = \mathbf{1}\left[\|g_{t,s}\|_2 \le C_t\right] - \frac{1}{2}. \tag{32}$$

The private estimate of the fraction below the threshold is

$$\widetilde{f}_t = \mathrm{clip}_{[0,1]}\left[\frac{1}{2} + \frac{\sum_{s \in \mathcal{S}_t} a_{t,s} + \sigma_b \zeta_t}{\rho N}\right], \qquad \zeta_t \sim \mathcal{N}(0,1), \tag{33}$$

and the next threshold is updated geometrically,

$$C_{t+1} = \mathrm{clip}_{[C_{\min}, C_{\max}]}\left[C_t \exp\left\{-\eta_C(\widetilde{f}_t - \gamma)\right\}\right], \tag{34}$$

where γ denotes the target unclipped fraction.

Following Theorem 1 of Andrew et al. (2021), the vector and clipping count queries can be represented as one joint Gaussian mechanism. Let σ denote the effective noise multiplier supplied to the privacy accountant and σ_b the standard deviation of the count query. The gradient noise multiplier is

$$\sigma_g = \left(\sigma^{-2} - (2\sigma_b)^{-2}\right)^{-1/2}, \qquad \sigma_b > \frac{\sigma}{2}. \tag{35}$$

Proposition 1: ATS leaves the outer privacy accounting unchanged. For fixed public values of ρ, σ, T, δ, and the adaptive clipping configuration, replacing uniform within-subject sampling by ATS does not change the subject-level privacy accounting.

Proof. Condition on the previously privatized training transcript. The current parameters θ_t and clipping threshold C_t are therefore fixed. Regardless of how ATS internally constructs $g_{t,s}$, Eq.11 ensures

$$\|\bar{g}_{t,s}\|_2 \le C_t, \tag{36}$$

while Eq.32 gives $|a_{t,s}| \le 1/2$.

Following Theorem 1 in Andrew et al. (2021), define the joint normalized contribution of one sampled subject as

$$u_{t,s} = \left(\frac{\bar{g}_{t,s}}{\sigma_g C_t}, \frac{a_{t,s}}{\sigma_b}\right). \tag{37}$$

Its norm is bounded by

$$\|u_{t,s}\|_2^2 \le \frac{1}{\sigma_g^2} + \frac{1}{4\sigma_b^2} \tag{38}$$

$$= \frac{1}{\sigma^2}, \tag{39}$$

where the final equality follows from Eq.35. Consequently, the joint release has the same sensitivity to noise ratio as a Gaussian mechanism with effective multiplier σ. Subject inclusion is still governed by the same Bernoulli probability ρ. Therefore the sampled Gaussian mechanism and its privacy accountant are unchanged by the internal choice of ATS versus uniform subject-window sampling. □

B APPENDIX

B.1 EXPERIMENTAL SETUP

Dataset and Implementation: We utilize MM-Fi Yang et al. (2023), a comprehensive multimodal dataset designed for non-intrusive 4D human sensing. The dataset comprises over 320K synchronized frames across five modalities: RGB, depth, LiDAR, mmWave Radar, and WiFi CSI over 40 subjects. It encompasses 27 action categories, consisting of 14 daily activities and 13 rehabilitation activities. We specifically use LiDAR, Radar and RGB modalities for our experiments. MM-Fi is particularly challenging for Radar human pose estimation due to its sparse Radar point clouds containing fewer than 100 points per frame compared to other benchmarks. To test the robustness of our approach, we adopt the dataset's three standard splits: random split, cross-subject split, and cross-environment split across all three protocols. We implement all our models on two NVIDIA RTX 4090 GPUs. We use PointTransformerV2 Wu et al. (2022) as backbone for LiDAR and Radar point clouds while DSTFormer Zhu et al. (2023) serves as backbone for RGB based model. Additionally, in the RGB setting we follow the two stage pipeline of generating 3D pose from 2D estimates. MM-Fi dataset consists of 2D keypoints in prior extracted from HRNet-48 backbone and we leverage this. We use Adam optimizer for our experiments with a batch size of 32. The total epochs are set to 50. The learning rate is set to 1e-3 which is then multiplied by 0.1 after 40 epochs. We report Mean Per Joint Position Error (MPJPE) and Procrustes-aligned MPJPE (PA-MPJPE), following the standard 3D HPE evaluation protocols.

Private models are evaluated only on MM-Fi protocol 3 under cross subject split with 32 training subjects treated as members and 8 held out subjects treated as non members. We train for $T = 2000$ private steps with subject sampling probability $\rho = 0.25, \delta = 10^{-5}, andK = 64$ ATS selected windows per sampled subject.

B.2 ADDITIONAL RESULTS OF FUSION METHOD ON PROTOCOL 1 AND 2

Model	Modality	Random		Cross-Environment		Cross-Subject	
		MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Protocol 1: Daily Activities							
Existing Methods							
Point Transformer Zhao et al. (2021); Fan et al. (2024)	Radar	73.09	55.60	88.28	68.79	75.96	58.70
DiffPose Gong et al. (2023); Fan et al. (2024)	Radar	73.44	56.83	86.35	66.87	70.31	54.12
mmDiff Fan et al. (2024)	Radar	65.26	50.35	82.73	63.87	65.62	50.23
MM-Fi Yang et al. (2023)	Radar	109.8	55.6	166.2	73.9	128.4	58.7
MM-Fi Yang et al. (2023)	RGB	263.3	80.0	276.6	83.0	267.7	81.2
MM-Fi Yang et al. (2023)	LiDAR	98.1	65.2	192.3	100.4	110.1	66.2
MM-Fi Yang et al. (2023)	RGB+LiDAR	94.3	63.9	159.6	84.5	109.4	65.8
Balanced Multimodal Qi et al. (2025)	Radar	102.89	52.21	–	–	–	–
Balanced Multimodal Qi et al. (2025)	RGB	63.61	35.75	–	–	–	–
Balanced Multimodal Qi et al. (2025)	LiDAR	66.95	45.70	–	–	–	–
Balanced Multimodal Qi et al. (2025)	RGB+LiDAR	52.93	34.96	–	–	–	–
Balanced Multimodal Qi et al. (2025)	RGB+Radar	63.04	35.74	–	–	–	–
Balanced Multimodal Qi et al. (2025)	LiDAR+Radar	66.85	45.11	–	–	–	–
Our Implementations							
PointTransformerV2	LiDAR	47.60	38.73	66.68	47.68	52.47	39.02
PointTransformerV2	Radar	61.65	49.42	83.27	60.55	62.14	46.12
DSTFormer	RGB	54.10	41.84	63.38	47.04	56.06	43.59
Fusion	RGB+LiDAR+Radar	**40.60**	**32.91**	**55.62**	**42.53**	**45.38**	**34.98**
Protocol 2: Rehabilitation Activities							
Existing Methods							
MM-Fi Yang et al. (2023)	RGB	291.6	83.6	301.5	85.3	304.3	82.6
MM-Fi Yang et al. (2023)	LiDAR	94.9	60.4	186.0	103.5	103.9	60.3
MM-Fi Yang et al. (2023)	Radar	124.3	55.3	168.0	73.0	142.2	57.4
MM-Fi Yang et al. (2023)	RGB+LiDAR	90.1	58.9	153.5	87.1	103.5	60.2
Our Implementations							
PointTransformerV2	LiDAR	42.76	32.97	61.01	46.43	41.79	32.43
PointTransformerV2	Radar	59.76	44.93	81.04	63.07	58.66	43.94
DSTFormer	RGB	41.59	32.70	54.93	41.38	42.51	33.02
Fusion	RGB+LiDAR+Radar	**38.64**	30.67	**53.00**	39.64	**41.39**	29.97

Table 3: Comparison on Protocol 1 (Daily Activities) and Protocol 2 (Rehabilitation Activities). Lower MPJPE and PA-MPJPE indicate better performance.

B.3 BIMODAL ABLATIONS

Protocol	Split	Bimodal Fusion	MPJPE ↓	PA-MPJPE ↓
Protocol 1: Daily Activities				
P1	Random	RGB+LiDAR	42.24	34.02
P1	Random	RGB+Radar	46.17	35.69
P1	Random	LiDAR+Radar	44.89	34.02
P1	Cross-Environment	RGB+LiDAR	55.68	42.68
P1	Cross-Environment	RGB+Radar	60.91	44.73
P1	Cross-Environment	LiDAR+Radar	58.66	45.04
P1	Cross-Subject	RGB+LiDAR	47.69	36.21
P1	Cross-Subject	RGB+Radar	48.60	36.44
P1	Cross-Subject	LiDAR+Radar	49.28	36.93
Protocol 2: Rehabilitation Activities				
P2	Random	RGB+LiDAR	41.66	31.96
P2	Random	RGB+Radar	42.80	32.68
P2	Random	LiDAR+Radar	42.11	33.41
P2	Cross-Environment	RGB+LiDAR	52.15	39.83
P2	Cross-Environment	RGB+Radar	52.39	39.88
P2	Cross-Environment	LiDAR+Radar	57.62	43.84
P2	Cross-Subject	RGB+LiDAR	42.73	32.26
P2	Cross-Subject	RGB+Radar	43.33	33.13
P2	Cross-Subject	LiDAR+Radar	43.80	34.01
Protocol 3: All Activities				
P3	Random	RGB+LiDAR	41.96	33.50
P3	Random	RGB+Radar	44.29	34.16
P3	Random	LiDAR+Radar	43.21	34.07
P3	Cross-Environment	RGB+LiDAR	54.47	41.37
P3	Cross-Environment	RGB+Radar	55.86	42.42
P3	Cross-Environment	LiDAR+Radar	67.03	49.71
P3	Cross-Subject	RGB+LiDAR	45.89	34.72
P3	Cross-Subject	RGB+Radar	47.48	35.74
P3	Cross-Subject	LiDAR+Radar	45.83	35.97

Table 4: Bimodal fusion ablation on MM-Fi across the three activity protocols and evaluation splits. Results are reported for RGB+LiDAR, RGB+Radar, and LiDAR+Radar combinations. MPJPE and PA-MPJPE are reported in millimeters; lower is better.

B.4 Comparison with BEV fusion

Split	Method	MPJPE (mm) ↓
Random	BEV Baseline	67.03
Random	Our Fusion	**43.21**
Cross-Environment	BEV Baseline	73.37
Cross-Environment	Our Fusion	**67.03**
Cross-Subject	BEV Baseline	56.44
Cross-Subject	Our Fusion	**45.83**

Table 5: Comparison with the BEV-based fusion baseline Ji et al. (2024) on MM-Fi Protocol 3 using the LiDAR+Radar bimodal setting across the random, cross-environment, and cross-subject splits. MPJPE is reported in millimeters; lower is better.